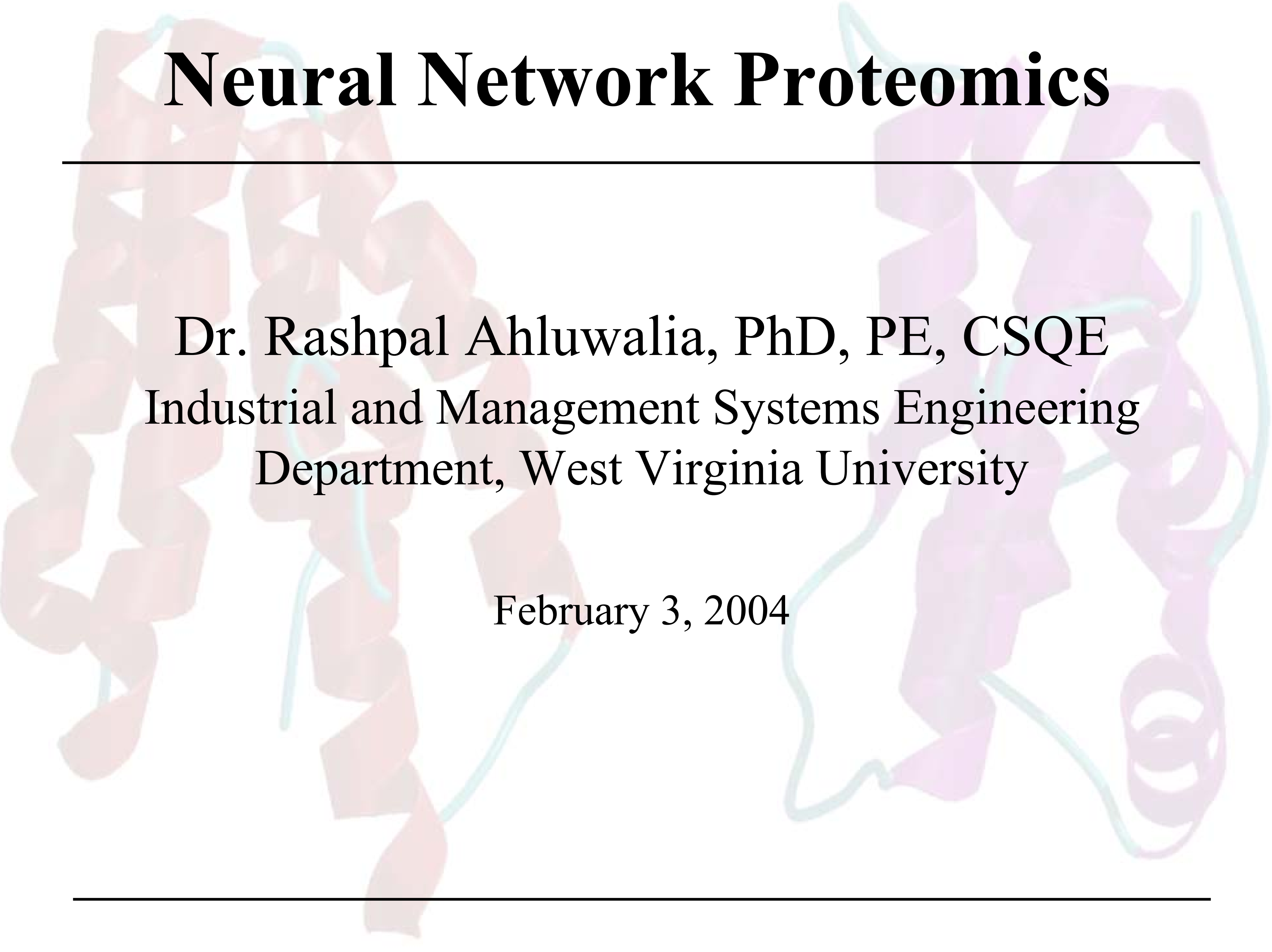
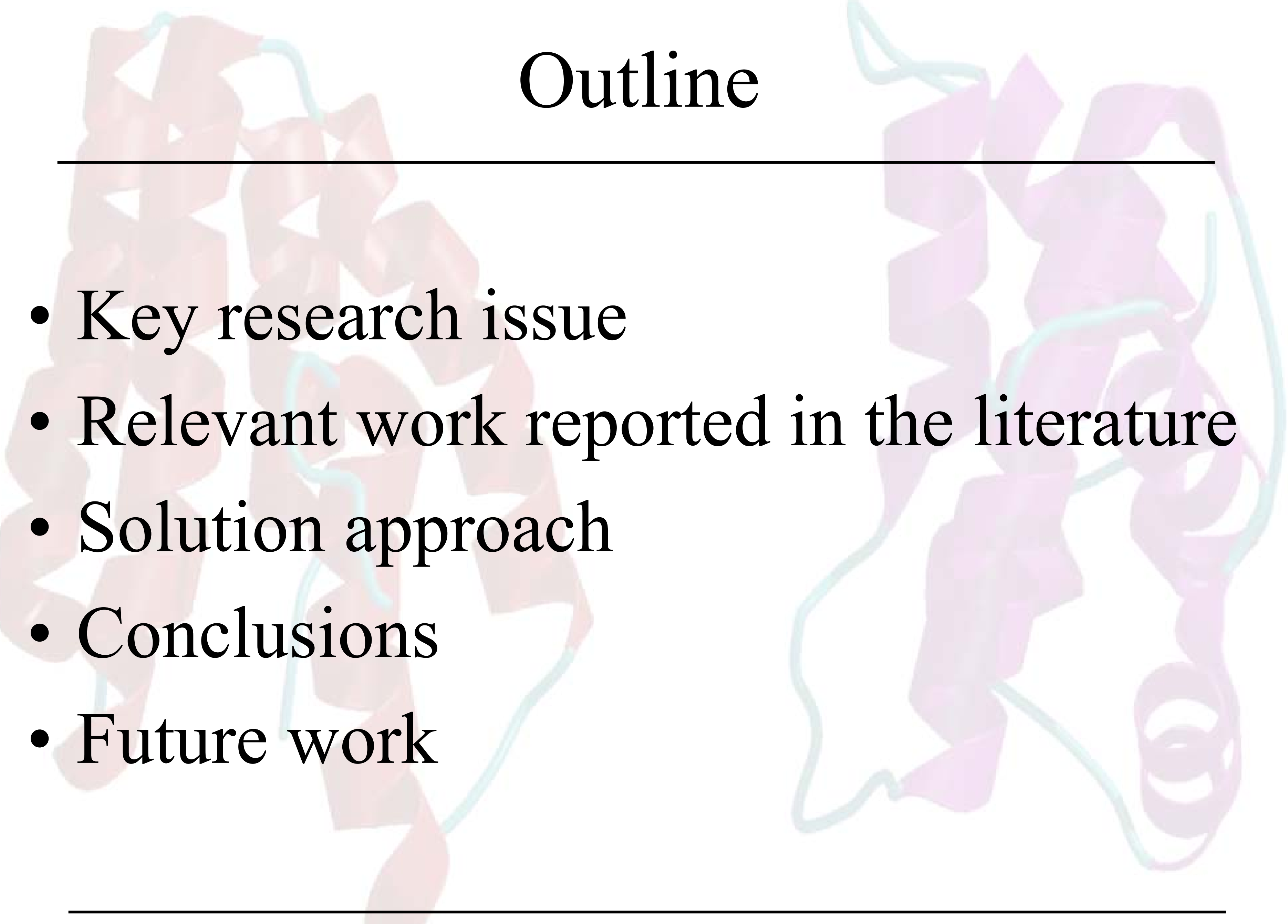
Neural Network Proteomics

Dr. Rashpal Ahluwalia, PhD, PE, CSQE
Industrial and Management Systems Engineering Department, West Virginia University

February 3, 2004





Outline

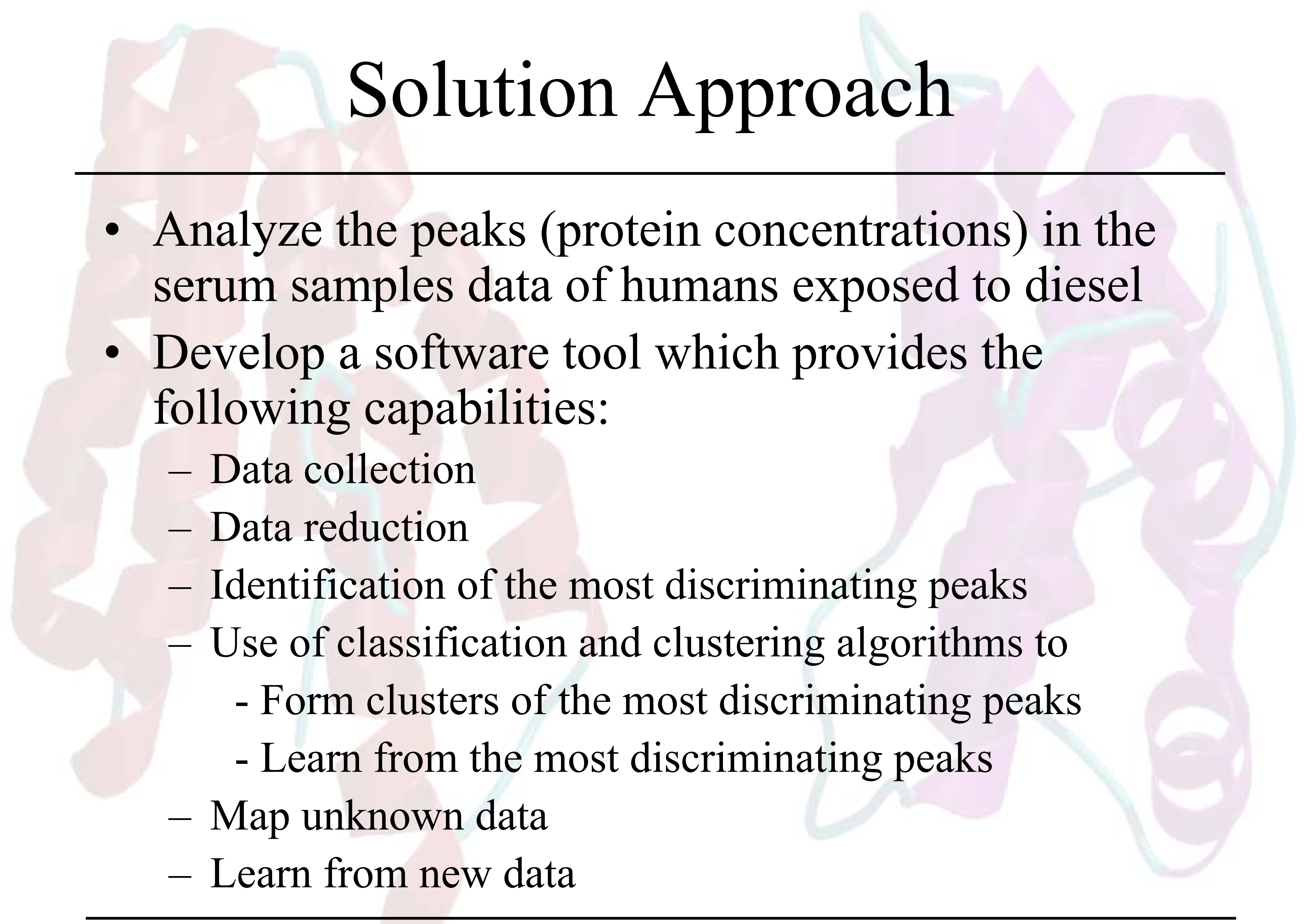
- Key research issue
- Relevant work reported in the literature
- Solution approach
- Conclusions
- Future work

Key Research Issue

Identify the most discriminating biomarkers of exposure and/or response to diesel exhaust

Relevant Work Reported in the Literature

- Petricoin EF III, Ardekani M, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA., "Use of Proteomic Patterns in Serum to identify Ovarian Cancer", Mechanisms of Disease, 2002, Vol. 359, pp: 572-577.
- Sorace JM, Zhan M., "A data review and re-assessment of ovarian cancer serum proteomic profiling", BMC Bioinformatics, 2003.
- Ball G, Mian S, Holding F, Allibone RO, Lowe J, Ali S, Li G, McCardle S, Ellis IO, Creaser C, and Rees RC., "An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumors and rapid identification of potential biomarkers", Bioinformatics, 2002, Vol. 18: Issue 3, pp: 393-404.
- Li J, Zhang Z, Rosezweig J, Wang YY, and Chan DW., "Proteomics and Bioinformatics Approaches for Identification of Serum Biomarkers to Detect Breast Cancer", Clinical Chemistry, 2002, Vol. 48, Issue 8, pp: 1296-1304.

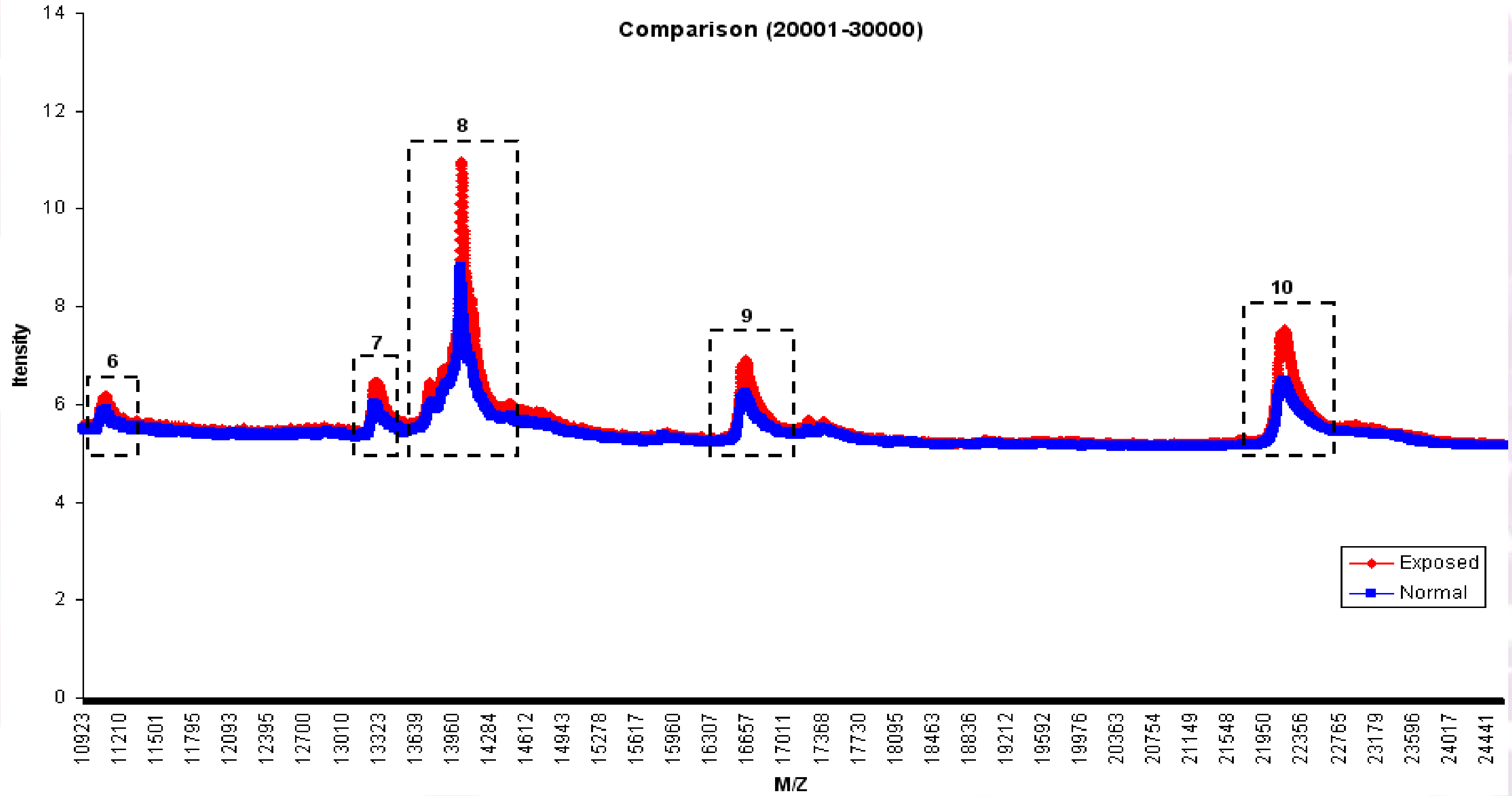


Solution Approach

- Analyze the peaks (protein concentrations) in the serum samples data of humans exposed to diesel
- Develop a software tool which provides the following capabilities:
  - Data collection
  - Data reduction
  - Identification of the most discriminating peaks
  - Use of classification and clustering algorithms to
    - Form clusters of the most discriminating peaks
    - Learn from the most discriminating peaks
  - Map unknown data
  - Learn from new data

Data Collection

Each spectrum had 60,000 data-points

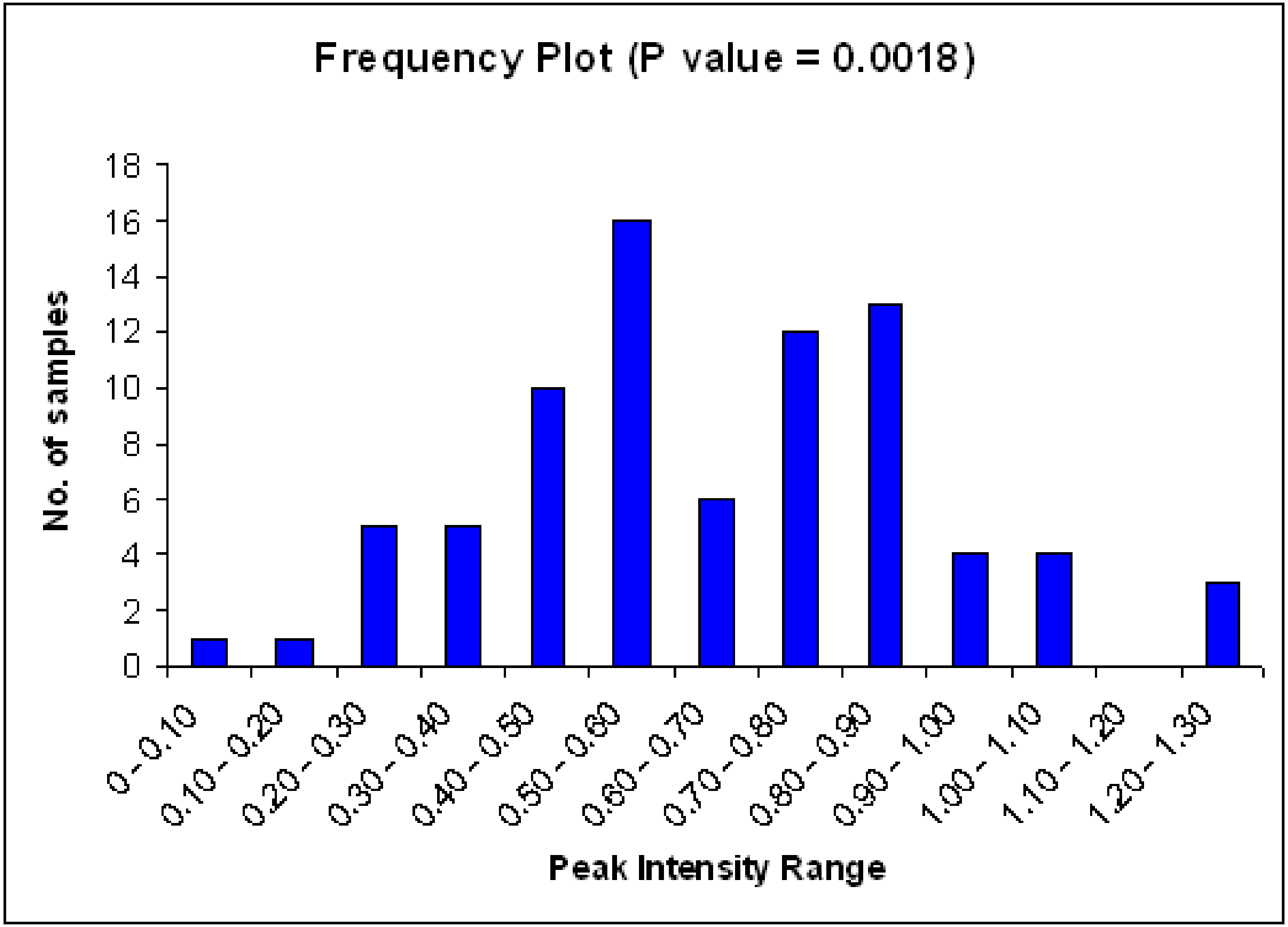


Snapshot of peaks in 10,000 points of the spectrum

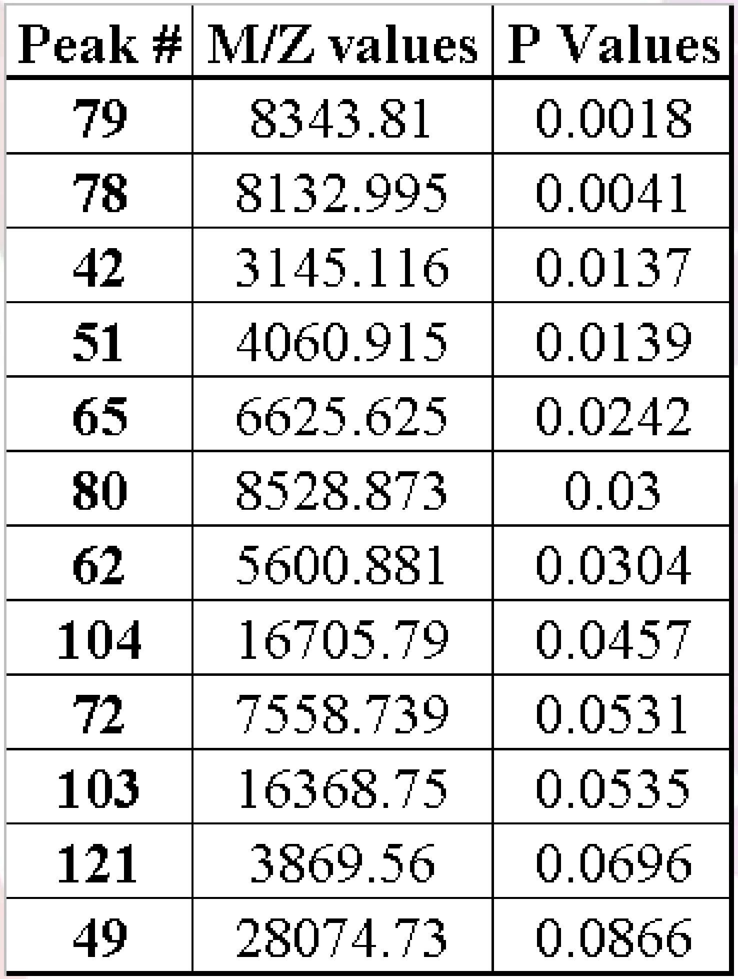
Data Reduction

- Data from 93 samples (34 – “low”, 13 – “medium” and 46 “high” exposures) were obtained from the Ciphergen software.
- Based on the 60,000 data points for each of the 93 samples, the Ciphergen software normalized the data, applied baseline correction and identified 132 peaks for each sample.
- The most differentiating peaks between the “high” and “low’ diesel exposures were obtained by performing a t-test for unequal variances and selecting peaks having p-values < 0.10.
- Intensities from the 93 samples for the 12 peaks were fed into the Predict software as training data.

Analysis of peaks using t-test



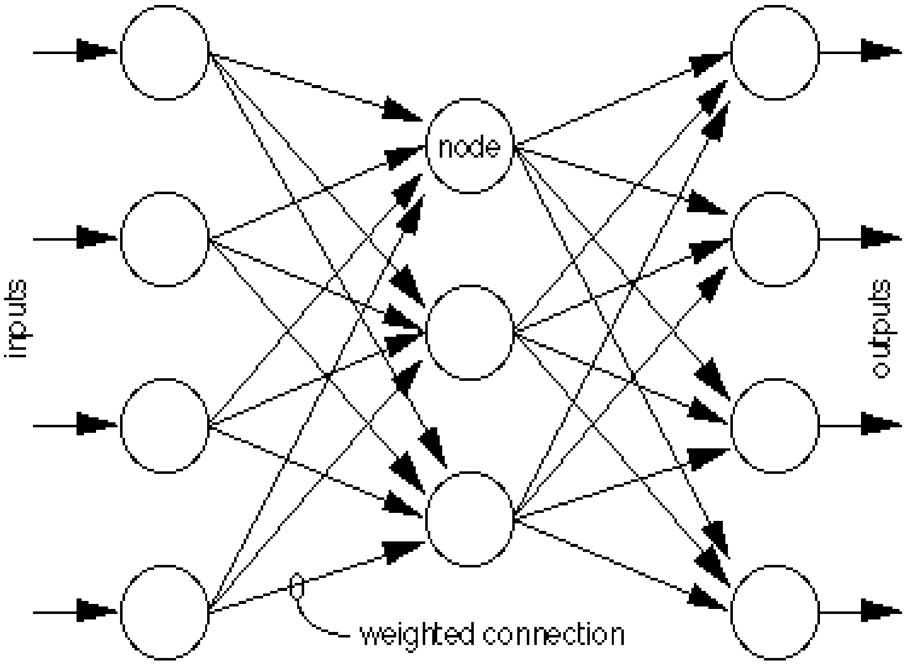
Identification of the most discriminating peaks (based on p-value in peak statistics)



Peak #	M/Z values	P Values
79	8343.81	0.0018
78	8132.995	0.0041
42	3145.116	0.0137
51	4060.915	0.0139
65	6625.625	0.0242
80	8528.873	0.03
62	5600.881	0.0304
104	16705.79	0.0457
72	7558.739	0.0531
103	16368.75	0.0535
121	3869.56	0.0696
49	28074.73	0.0866

The Backpropagation Algorithm (BPA)

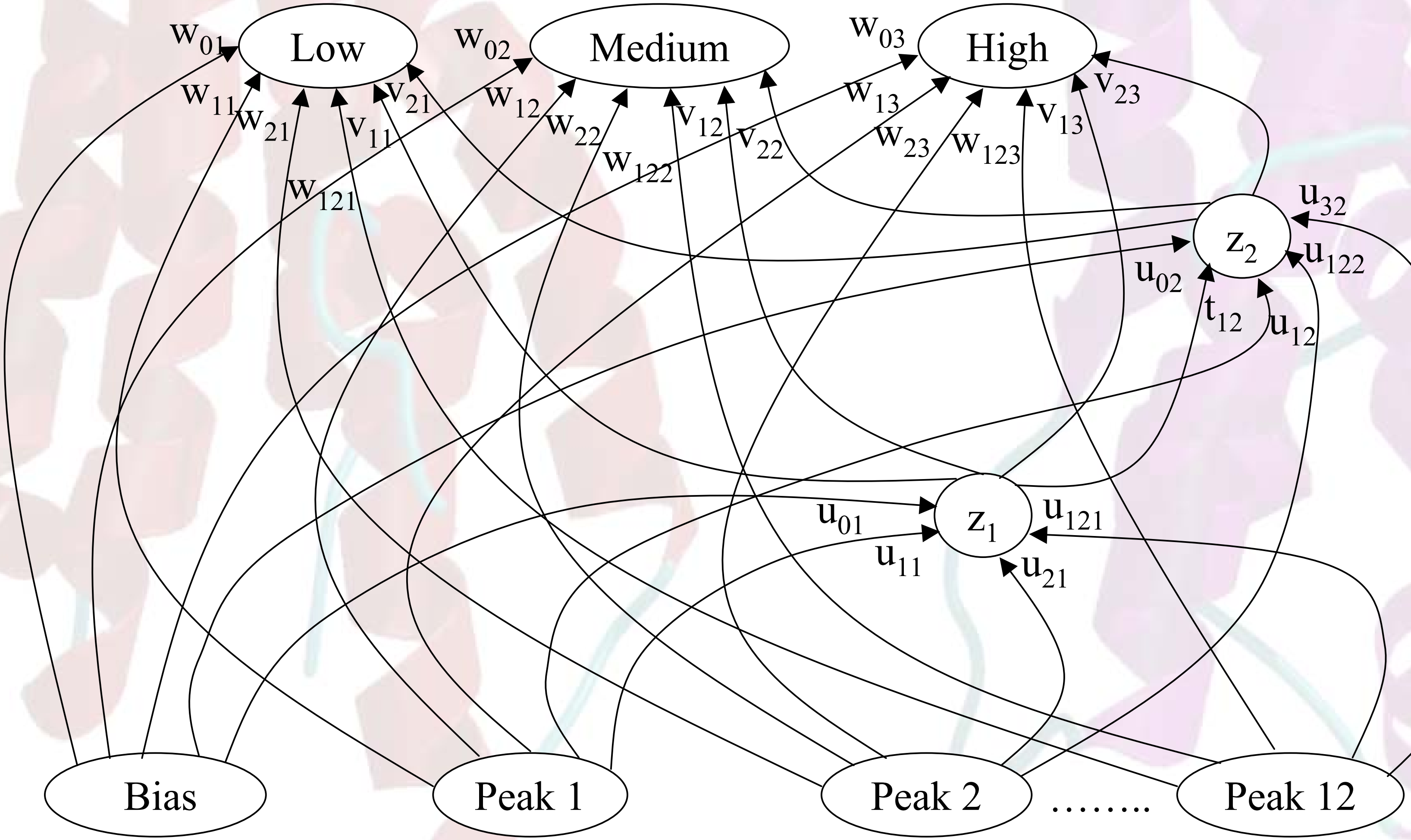
- This algorithm is a standard error backpropagation algorithm that uses error propagation to identify patterns
- The algorithm learns from its errors and it falls under the category of supervised learning algorithms
- Typically it has three phases:
  - Feed forward phase of input training pattern
  - Back propagation of associated error
  - Adjustment of weights



The Cascade Correlation Algorithm (CCA)

- Dynamically adds hidden units (only the minimum number necessary to achieve the specified error).
- A two-step weight training process ensures that only one layer of weights is being trained at any time.
- The network is trained until no further improvement is obtained (error of the output unit over all training patterns remains unchanged)
- The example shows 12 peak intensities as inputs and the corresponding “low”, “medium” and “high” exposures as outputs.

Entire Network of CCA

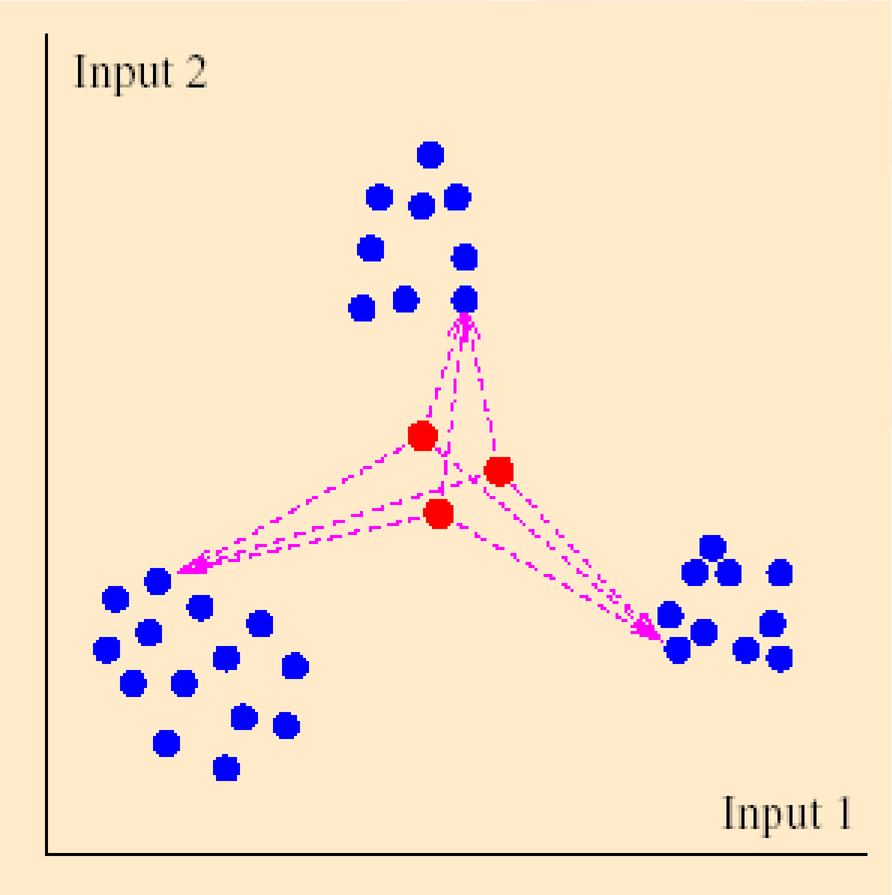


The SOM Algorithm

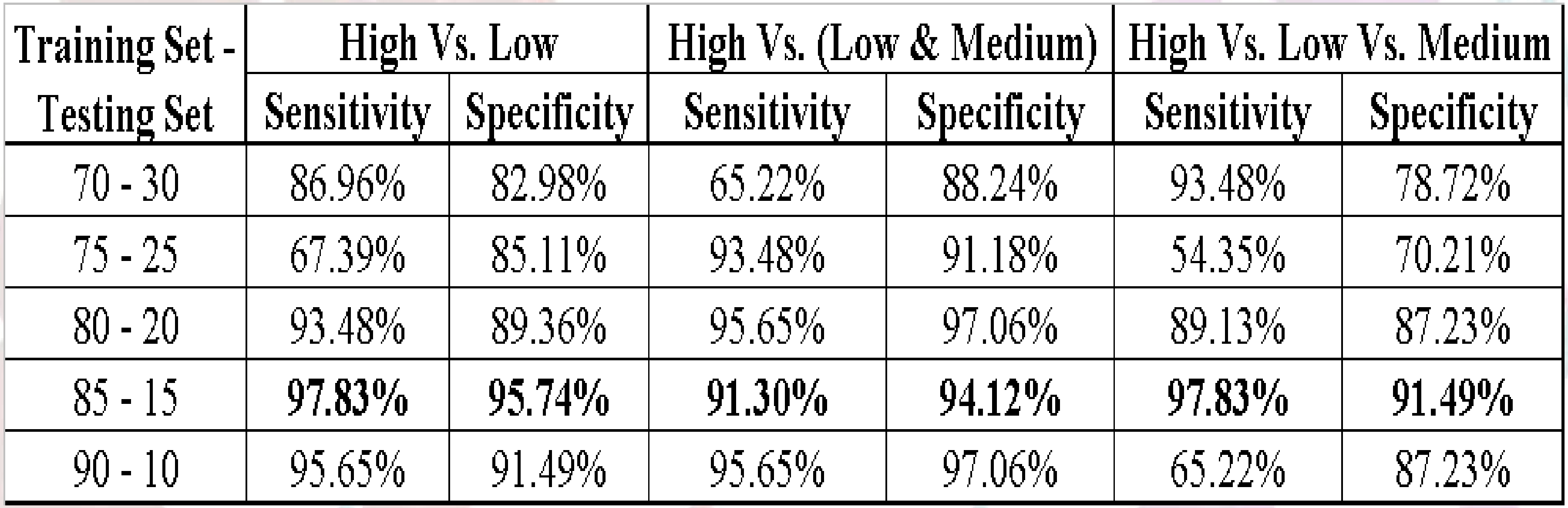
- The Self Organizing Maps – also called the "Winner Take All" algorithm
- Algorithm clusters data based on Euclidean Distance
- Algorithm falls under the category of Unsupervised Algorithms

Working of SOM

- A base level of activation is calculated based on the distance between the weight vectors and the input vectors being examined
- A "Processing Element" with the highest level of activity is the "Winner"
- Once the "Winner" is decided, the activation levels of all other units are squashed to zero. Allows the winner to learn only from the current input pattern



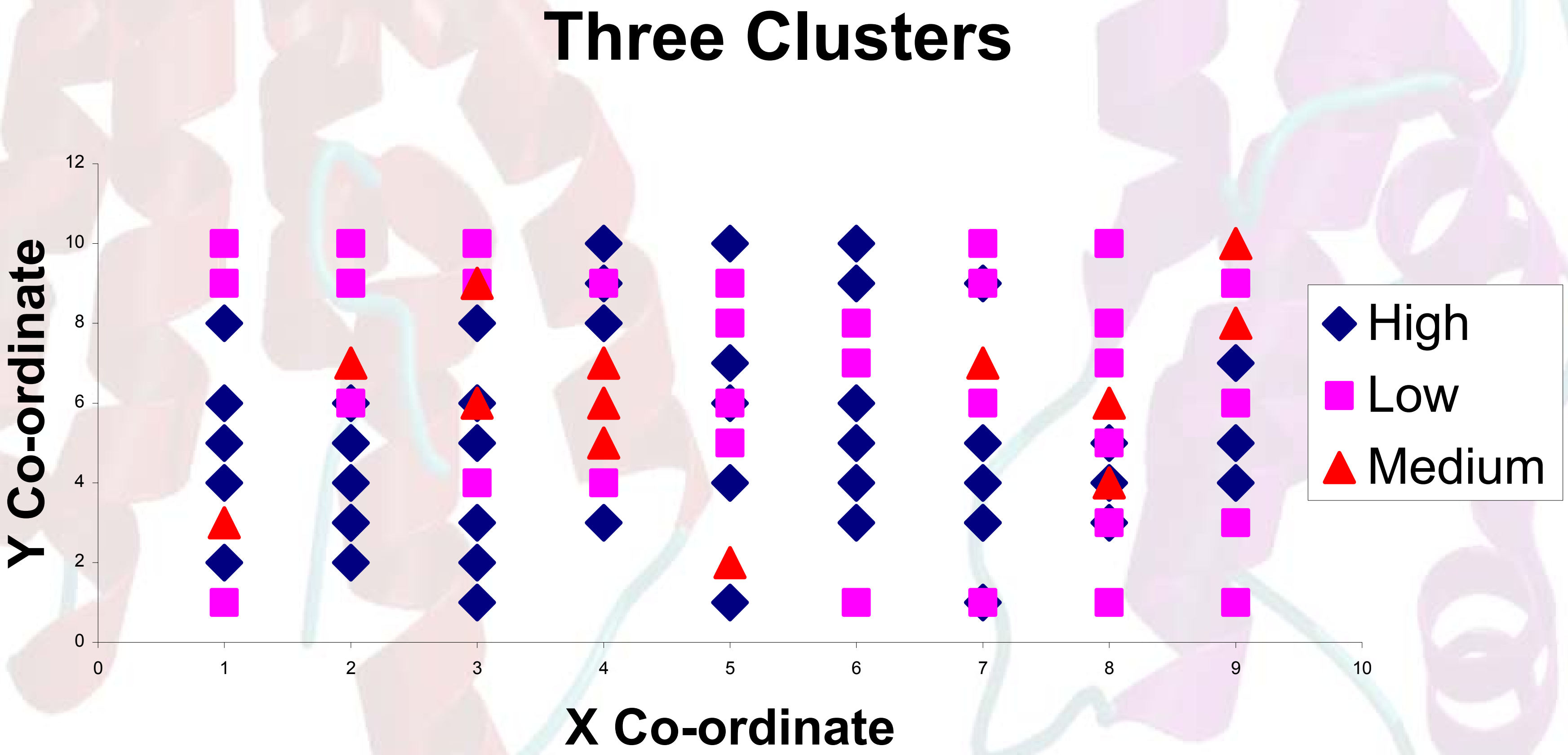
Analysis of Results (Classification Algorithm)



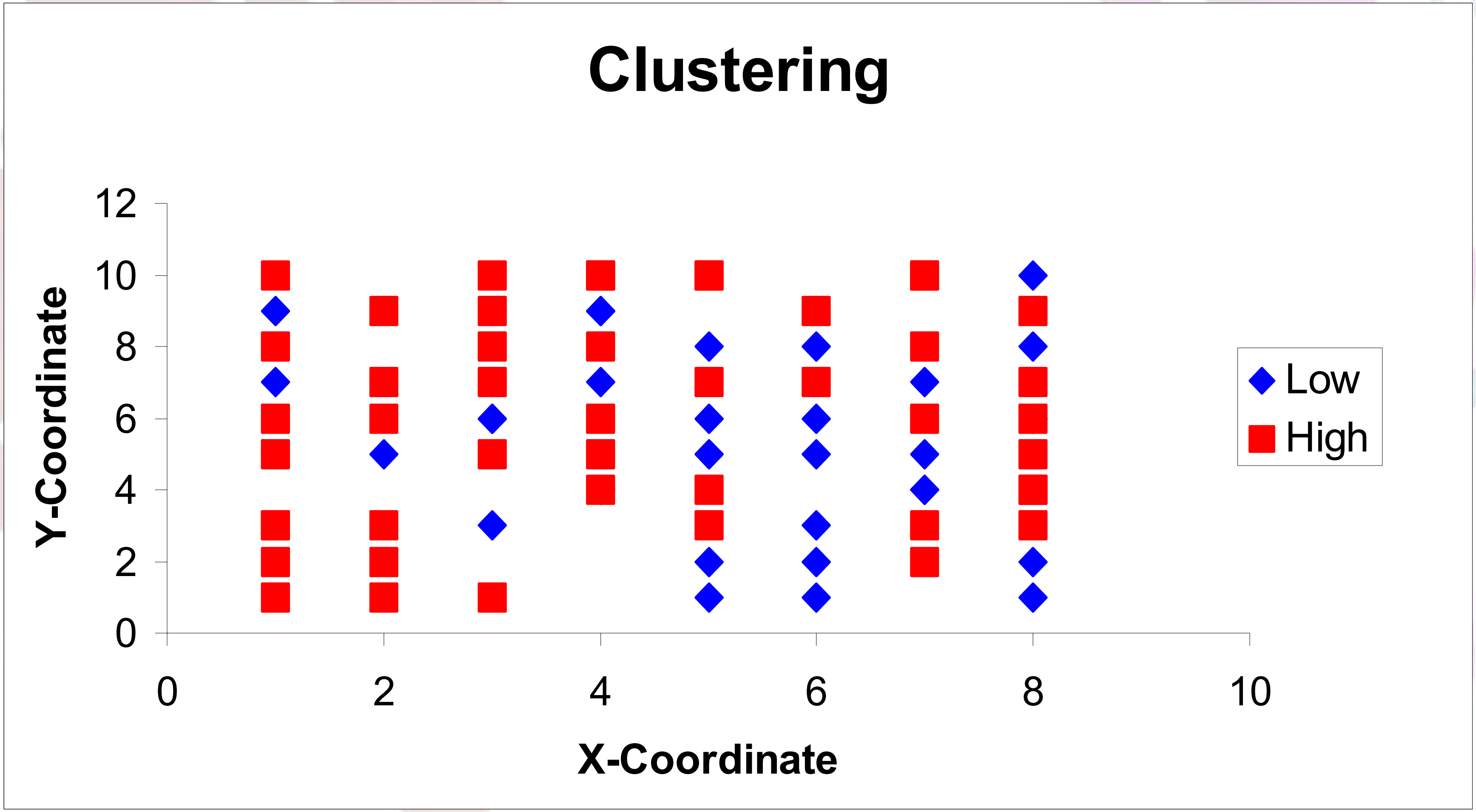
Training Set - Testing Set	High Vs. Low		High Vs. (Low & Medium)		High Vs. Low Vs. Medium	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
70 - 30	86.96%	82.98%	65.22%	88.24%	93.48%	78.72%
75 - 25	67.39%	85.11%	93.48%	91.18%	54.35%	70.21%
80 - 20	93.48%	89.36%	95.65%	97.06%	89.13%	87.23%
85 - 15	**97.83%**	**95.74%**	**91.30%**	**94.12%**	**97.83%**	**91.49%**
90 - 10	95.65%	91.49%	95.65%	97.06%	65.22%	87.23%

The "Predict" software provided the best results when 85% of the data was used in the training set and the remaining 15% of the data was used as the testing set.

Analysis of Results (Clustering Algorithm)



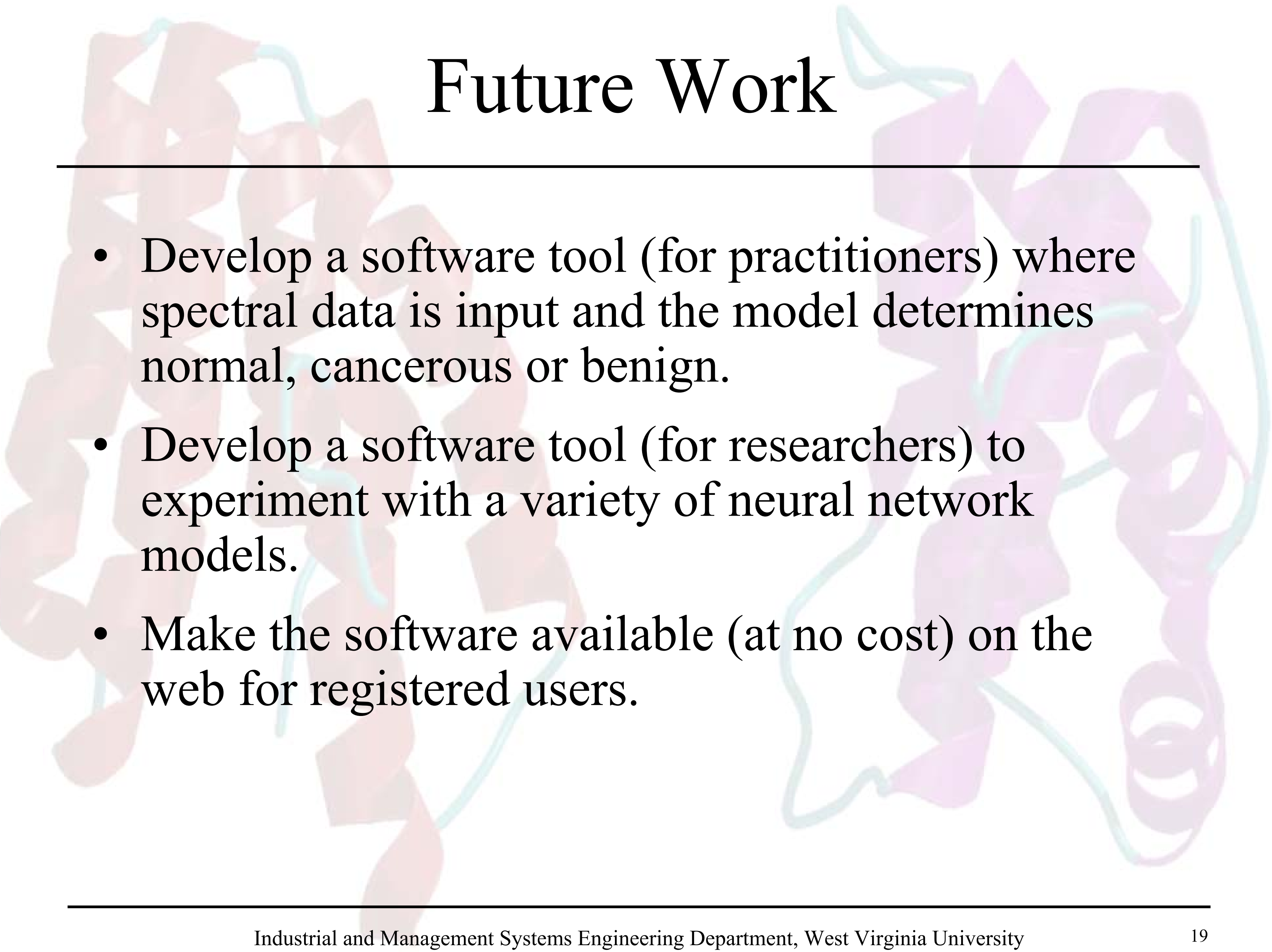
Analysis of Results (Clustering Algorithm)



Conclusions

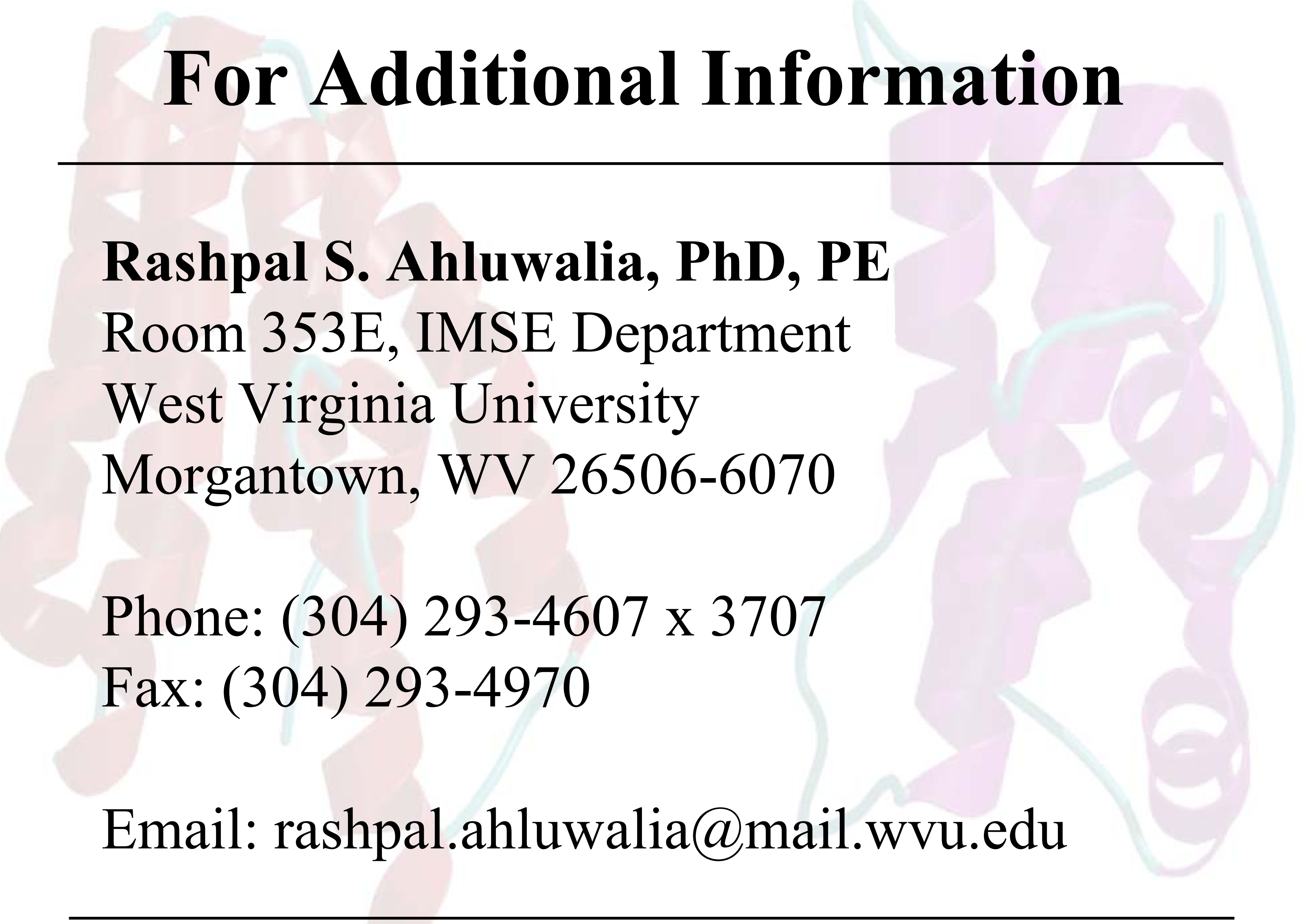
- The neural network algorithms showed high levels of sensitivities and specificities.
- Adapt the neural network algorithms for biological data.
- Develop algorithm(s) to automatically identify the most discriminating peaks.
- Develop algorithm(s) to cluster cases based on a variety of criteria. Identify relationships between different clusters formed for different criteria.
- Identify the most robust tools and techniques, both statistical and neural network, for proteomic data.

Future Work



- Develop a software tool (for practitioners) where spectral data is input and the model determines normal, cancerous or benign.
- Develop a software tool (for researchers) to experiment with a variety of neural network models.
- Make the software available (at no cost) on the web for registered users.

For Additional Information



Rashpal S. Ahluwalia, PhD, PE
Room 353E, IMSE Department
West Virginia University
Morgantown, WV 26506-6070

Phone: (304) 293-4607 x 3707
Fax: (304) 293-4970

Email: rashpal.ahluwalia@mail.wvu.edu